An Akaike information criterion for multiple event mixture cure models

Lore Dirick[1], **Gerda Claeskens**[1], **Bart Baesens**[2]

[1] ORSTAT and Leuven Statistics Research Center; [2] LIRIS

KU Leuven, Faculty of Economics and Business

Naamsestraat 69, 3000 Leuven, Belgium.

Lore.Dirick@kuleuven.be; Gerda.Claeskens@kuleuven.be; Bart.Baesens@kuleuven.be

May 21, 2014

Abstract

We derive the proper form of the Akaike information criterion for variable selection for mixture cure models, which are often fit via the expectation-maximization algorithm. Separate covariate sets may be used in the mixture components. The selection criteria are applicable to survival models for right-censored data with multiple competing risks and allow for the presence of an insusceptible group. The method is illustrated on credit loan data, with pre-payment and default as events and maturity as the insusceptible case and is used in a simulation study.

Keywords: Akaike information criterion, Competing risks, EM-algorithm, Mixture cure model, Model selection.

1 Introduction

The topic of credit risk modeling has now become more important than ever before. The introduction of compliance guidelines such as Basel II and Basel III have a huge impact

We acknowledge the support of the Fund for Scientific Research Flanders, KU Leuven grant GOA/12/14 and of the IAP Research Network P7/06 of the Belgian Science Policy. The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Hercules Foundation and the Flemish Government - department EWI.

on the strategies of financial institutions nowadays. The Basel Accords aim at quantifying the minimum amount of buffer capital so as to provide a safety cushion against unexpected losses (Van Gestel and Baesens, 2008). A key credit risk parameter is the probability of default (PD) measuring the likelihood of an obligor to run into arrears on his/her credit obligation.

PD models are typically constructed using classification techniques such as logistic regression (Baesens et al., 2003). However, the timing when customers default is perhaps of even more interest to analyse since it can provide the bank with the ability to compute the profitability over a customer's lifetime and perform profit scoring. The problem statement of analysing when customers default is commonly referred to as survival analysis (see, e.g., Bellotti and Crook, 2009). It is the purpose of this paper to provide a valid model selection criterion for variable selection inside such survival models, specifically applied to credit risk modelling, with as particular characteristics allowing for defaults, maturity and early repayments in a mixture cure rate model and allowing for right-censored data.

In this paper we deal with right-censored failure times in a mixture model context. This implies that there are two sources of incompleteness: (i) the right-censoring causes some of the event times to remain unobserved, it is only known that the event of interest did not yet take place, and (ii) not for all observations it is known to which component of the mixture model they belong; in fact, only when an observation is uncensored, we have this information. For this type of cure rate models no information criteria have yet been derived.

For incomplete and partially observed data, Cavanaugh and Shumway (1998) derive a version of the Akaike information criterion (AIC Akaike, 1973) that makes use of the expected complete data log-likelihood, rather than the observed log-likelihood. They coined the name AICcd to this criterion. The use of the likelihood for the observed cases is discouraged since a comparison of this 'model' likelihood to a 'true' likelihood

for the observed cases only is rarely of interest. By working with the complete data log-likelihood, and considering the Kullback-Leibler distance between the model and true data generating process for the complete data, the AICcd is able to select models, taking unobserved and latent variables into account. The method uses directly the output of the expectation-maximization algorithm (EM). We explain its definition and use below. For a comprehensive explanation of the AIC, see Claeskens and Hjort (2008, Chap. 2).

Similar variations on the AIC are studied by Claeskens and Consentino (2008), who use the output of an EM algorithm to define variable selection methods for models with missing covariate data in a linear regression setting and by Ibrahim et al. (2008) for missing data variable selection in generalized linear models.

For the case of right-censored data (not in a mixture), Liang and Zou (2008) work with an accelerated life time model and propose for that model a finite sample correction to the standard AIC, motivated from an exponential model with constant censoring. For parametric survival models Suzukawa et al. (2001) derive a version of the AIC taking the censoring into account, though require a non-standard estimation method for practical use. Fan and Li (2002) used a smoothly clipped absolute deviation penalty for the semiparametric Cox proportional hazard models, Hjort and Claeskens (2006) derived a focussed information criterion, while Xu et al. (2009) define an AIC based on the profile likelihood for proportional hazard mixed models, see also Donohue et al. (2011) for a related model selection approach. None of these papers made use of the EM algorithm to define the variable selection criterion, and neither did they consider mixture models.

In Section 2 we first consider the Akaike information criterion for the case of a mixture cure model with one event of interest and a group non-susceptible to this event. In Section 3 we extend the applicability of the AIC to the model recently proposed by Watkins et al. (2013) that provides a simultaneous modeling of multiple event times, potentially right censored, in the presence of a nonsusceptible group. While parametric

survival models can be used as in the approach of Watkins et al. (2013), in this paper we use the semiparametric Cox proportional hazard model for the susceptible part(s) of the mixture model and we use logistic regression for the incidence part. Simulation results are given in Section 4 and an application to credit loan data is presented in Section 5.

2 The mixture cure model for a single event

Mixture cure models were motivated by the existence of a subgroup of long-term survivors, or 'immunes' in a medical context. This subgroup, with survival probabilities set equal to one, is incorporated in a model through a mixture distribution where a logistic regression model provides a mixing proportion of the 'non-susceptible' cases and where a survival model describes the cases susceptible to the event of interest. Such models were introduced by Farewell (1982) in a parametric version, and later generalized to a semi-parametric mixture model combining logistic regression and Cox proportional hazards regression by Kuk and Chen (1992), see also Sy and Taylor (2000). Recently, Cai et al. (2012) introduced the R-package `smcure` to estimate such semi-parametric mixture models.

Tong et al. (2012) use a mixture cure approach to analyze the credit risk of a specific customer, where the event of interest is the time of default when customers stop paying back their loans. This setting is characterized and distinguishes itself from typical medical settings by a heavy right-censoring, since most customers do not default. A relatively large group of non-susceptible cases is expected to be present. Part of the explanation of this high percentage of censoring is that both prepayments and maturity (loan completely paid back on time) are considered censored for default. For a separate analysis of default and prepayment, see, e.g., Stepanova and Thomas (2002).

2.1 Model notation

We denote the 'true' event time by U and the censoring time by C. We assume independence between event times and censoring times. Denote by Y a binary random variable where $Y = 1$ expresses susceptibility to the event of interest and $Y = 0$ indicates that the event will never happen. When $U > C$, the event is right-censored; the observed event time $T = \min(U, C)$. Let the indicator $\delta = I(U \leq C)$, thus $\delta = 1$ indicates non-censored observations. The combination of values for Y and δ generates three different states:

(1) $Y = 1$ and $\delta = 1$: uncensored and susceptible, so the event takes place during the observation period of the data;

(2) $Y = 1$ and $\delta = 0$: censored and susceptible, no event during observation period, but it will eventually take place;

(3) $Y = 0$ and $\delta = 0$: censored and non-susceptible, no event is observed, nor will it take place in the future.

Note that values for T and δ are fully observed while Y is only observed when $\delta = 1$ and is latent otherwise. Similarly, we do not observe U when $\delta = 0$. The sample information consists of values (T_i, δ_i), for $i = 1, \ldots, n$, together with some covariate information.

The incidence model component uses logistic regression to model $P(Y = 1) = \pi(\boldsymbol{z}; \boldsymbol{b})$ with $\text{logit}\{\pi(\boldsymbol{z}; \boldsymbol{b})\} = \boldsymbol{z}'\boldsymbol{b}$ for a r-vector of covariates $\boldsymbol{z} = (z_1, \ldots, z_r)'$. For the latency model, a semiparametric Cox proportional hazard regression model is used such that the conditional survival probability at time t is modeled as

$$S(t \mid Y = 1, \boldsymbol{x}; \boldsymbol{\beta}) \;\; = \;\; \exp\left(- \exp(\boldsymbol{x}^T \boldsymbol{\beta}) \int_0^t h_0(u \mid Y = 1) du \right),$$

with h_0 the unspecified baseline hazard function and $\boldsymbol{x}$ a q-vector of covariates $\boldsymbol{x} = (x_1, \ldots, x_q)'$, which may or may not contain the same components as $\boldsymbol{z}$. This yields the unconditional survival function

$$S(t, \boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\beta}, \boldsymbol{b}) = \pi(\boldsymbol{z}; \boldsymbol{b}) S(t \mid Y = 1, \boldsymbol{x}; \boldsymbol{\beta}) + 1 - \pi(\boldsymbol{z}; \boldsymbol{b}), \tag{1}$$

and the observed likelihood

$$\begin{aligned} L_{\text{obs}}(\boldsymbol{b}, \boldsymbol{\beta}) &= \prod_{i=1}^{n} \{\pi(\boldsymbol{z}_i; \boldsymbol{b}) f(t_i \mid Y_i = 1, \boldsymbol{x}_i)\}^{\delta_i} \\ &\quad \times \{(1 - \pi(\boldsymbol{z}_i; \boldsymbol{b})) + \pi(\boldsymbol{z}_i; \boldsymbol{b}) S(t_i \mid Y_i = 1, \boldsymbol{x}_i; \boldsymbol{\beta})\}^{1-\delta_i}. \end{aligned} \tag{2}$$

The complete likelihood, given full information on Y, can be expressed as:

$$L_{\text{complete}}(\boldsymbol{b}, \boldsymbol{\beta}) = \big(1 - \pi(\boldsymbol{z}_i; \boldsymbol{b})\big)^{(1-Y_i)} (\boldsymbol{z}_i; \boldsymbol{b})^{Y_i} h(t_i \mid Y_i = 1, \boldsymbol{x}_i; \boldsymbol{\beta})^{\delta_i, Y_i} S(t_i \mid Y_i = 1, \boldsymbol{x}_i; \boldsymbol{\beta})^{Y_i}$$

2.2 The Akaike information criterion for single event models

For estimation of mixture cure models, Cai et al. (2012) explain the use of the expectation-maximization (EM) algorithm to deal with the latent Y values. If $Y = Y^*$ would be observed for all cases, the log-likelihood for the data triplets (T_i, δ_i, Y_i) could be used in the AIC to lead to the (infeasible)

$$\text{AIC}_{\text{infeasible}} = -2 \log L_{T,\delta,Y}(\widehat{\boldsymbol{\Theta}}; T_i, \delta_i, Y_i^*) + 2d, \tag{3}$$

where d counts the number of parameters in the model, and $\widehat{\boldsymbol{\Theta}}$ is the maximum likelihood estimator of the parameter vector $\boldsymbol{\Theta}$.

The AIC estimates the expected value of the Kullback-Leibler discrepancy between the model and the unknown true data-generating process, without having to know this true process. In the general case with random variables $\boldsymbol{R} = (R_1, \ldots, R_n)$, a model $f(r; \boldsymbol{\Theta})$ and the density of the true data-generating process $g(r)$, the Kullback-Leibler (KL) discrepancy is given by $\text{KL}\{g, f(\cdot; \boldsymbol{\Theta})\} = \text{E}_g\Big\{ \log \dfrac{g(\boldsymbol{R})}{f(\boldsymbol{R}; \boldsymbol{\Theta})} \Big\}$, where the subscript g reminds of using the true density function g to compute the expectation. Since $\text{E}_g[\log g(\boldsymbol{R})]$ does not vary when searching through several candidate models, minimizing $\text{KL}\{g, f(\cdot; \boldsymbol{\Theta})\}$ over different models is equivalent with minimizing the quantity $D_{\boldsymbol{R}}(\boldsymbol{\Theta}) = \text{E}_g\{-2 \log f(\boldsymbol{R}; \boldsymbol{\Theta})\}$, where the expectation is computed using the true density function of the data. In our

notation $\boldsymbol{R} = (T, \delta, Y)$, which can be split in an observed vector $\boldsymbol{R}_{\text{obs}} = (T, \delta)$ and a "missing" part $\boldsymbol{R}_{\text{mis}} = Y$ indicating that Y is not always observed.

By rewriting the true joint density of the vector $\boldsymbol{R}$ as $g(r) = g_{Y|T,\delta}(y|t,\delta) \cdot g_{T,\delta}(t,\delta)$, it follows that $D_{\boldsymbol{R}}(\boldsymbol{\Theta}) = \mathrm{E}_{[T,\delta]}[-2Q^*(\boldsymbol{\Theta})]$, with the expected complete data log-likelihood $Q^*(\boldsymbol{\Theta}) = \mathrm{E}[\log f_{T,\delta,Y}(T, \delta, Y; \boldsymbol{\Theta})|T, \delta]$. The AIC procedure estimates $\mathrm{E}[D_{\boldsymbol{R}}(\widehat{\boldsymbol{\Theta}})]$ using the sample information.

Define $\boldsymbol{\Theta}_0$ as the least false parameter value that minimizes the KL discrepancy between the model density $f(\cdot; \boldsymbol{\Theta})$ and the true density g, $\boldsymbol{\Theta}_0 = \arg\min_{\boldsymbol{\Theta}} \mathrm{KL}\{g, f(\cdot; \boldsymbol{\Theta})\}$. As used in the EM algorithm, for two values, $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$ of the parameter vector $\boldsymbol{\Theta} = (\boldsymbol{b}, \boldsymbol{\beta})$ the expected complete-data log likelihood applied to our problem can be estimated by, see also Cai et al. (2012),

$$Q(\boldsymbol{\Theta}_2 \mid \boldsymbol{\Theta}_1) \;=\; \sum_{i=1}^{n} \int \log f_{T,\delta,Y}(T_i, \delta_i, Y = y; \boldsymbol{\Theta}_2) f_{Y|T,\delta}(y \mid T_i, \delta_i; \boldsymbol{\Theta}_1) dy \qquad (4)$$

where $f_{T,\delta,Y}$ denotes the joint density of the triplet (T, δ, Y) and where $f_{Y|T,\delta}$ is the probability mass function of Y conditional on (T, δ). Denote the first partial derivative $\dot{Q}(\boldsymbol{\Theta}_2 \mid \boldsymbol{\Theta}_1) = \frac{\partial}{\partial \boldsymbol{\Theta}_2} Q(\boldsymbol{\Theta}_2 \mid \boldsymbol{\Theta}_1)$ and the second partial derivative $\ddot{Q}(\boldsymbol{\Theta}_2 \mid \boldsymbol{\Theta}_1) = \frac{\partial}{\partial \boldsymbol{\Theta}_2 \partial \boldsymbol{\Theta}_2'} Q(\boldsymbol{\Theta}_2 \mid \boldsymbol{\Theta}_1)$. The EM approach proceeds by maximizing $Q(\boldsymbol{\Theta}_2 \mid \boldsymbol{\Theta}_1)$ over $\boldsymbol{\Theta}_2$, and by replacing the current $\boldsymbol{\Theta}_1$ by the maximizer. These steps are iterated until convergence. The resulting value of $\boldsymbol{\Theta}$ is denoted by $\widehat{\boldsymbol{\Theta}}$.

In the context of missing data, Claeskens and Consentino (2008) prove in their Theorem 1 that for a model density f that is two times continuously differentiable with respect to $\boldsymbol{\Theta}$, and which has a bounded expectation of the second derivative in a neighborhood of $\boldsymbol{\Theta}_0$, which belongs to the interior of a compact parameter space, if $n(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0)'(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0)$ is uniformly integrable, with the prime denoting a transpose, then

$$\mathrm{E}[D_{\boldsymbol{R}}(\widehat{\boldsymbol{\Theta}}) - Q(\widehat{\boldsymbol{\Theta}}|\widehat{\boldsymbol{\Theta}})]/n = \mathrm{trace}\{\boldsymbol{I}^{-1}(\boldsymbol{\Theta}_0) \cdot \boldsymbol{J}(\boldsymbol{\Theta}_0)\}/n + o(1/n),$$

where $\boldsymbol{I}(\boldsymbol{\Theta}) = \mathrm{E}\{-\ddot{Q}(\boldsymbol{\Theta} \mid \boldsymbol{\Theta})/n\}$, and $\boldsymbol{J}(\boldsymbol{\Theta}) = \mathrm{Var}\{\dot{Q}(\boldsymbol{\Theta} \mid \boldsymbol{\Theta})\}/n$.

Following Cavanaugh and Shumway (1998), by first taking a Taylor series expansion of $\dot{Q}(\boldsymbol{\Theta}_0 \mid \widehat{\boldsymbol{\Theta}})$ around $\widehat{\boldsymbol{\Theta}}$, leads to estimate $\boldsymbol{J}(\boldsymbol{\Theta}_0)$ by $\boldsymbol{I}(\widehat{\boldsymbol{\Theta}})\boldsymbol{I}_o^{-1}(\widehat{\boldsymbol{\Theta}})\boldsymbol{I}(\widehat{\boldsymbol{\Theta}})$, and further to estimate $\boldsymbol{I}(\boldsymbol{\Theta}_0)$ by $\boldsymbol{I}_{oc}(\widehat{\boldsymbol{\Theta}})$, where

$$\boldsymbol{I}_{oc}(\widehat{\boldsymbol{\Theta}}) = -n^{-1}\frac{\partial^2 Q(\widehat{\boldsymbol{\Theta}} \mid \widehat{\boldsymbol{\Theta}})}{\partial\boldsymbol{\Theta}\cdot\partial\boldsymbol{\Theta}'}, \qquad \boldsymbol{I}_o(\widehat{\boldsymbol{\Theta}}) = -n^{-1}\sum_{i=1}^{n}\frac{\partial^2 \log f_{T,\delta}(T_i,\delta_i;\widehat{\boldsymbol{\Theta}})}{\partial\boldsymbol{\Theta}\cdot\partial\boldsymbol{\Theta}'}.$$

This leads us to define the complete data AIC by

$$\text{AICcd} = -2Q(\widehat{\boldsymbol{\Theta}} \mid \widehat{\boldsymbol{\Theta}}) + 2\ \text{trace}\{\boldsymbol{I}_{oc}(\widehat{\boldsymbol{\Theta}})\cdot\boldsymbol{I}_o^{-1}(\widehat{\boldsymbol{\Theta}})\}, \tag{5}$$

Note that this derivation has relaxed the strong assumption of Cavanaugh and Shumway (1998) to have the model correctly specified, that is, they assumed that $g(r) = f(r;\boldsymbol{\Theta}_0)$. By working with least false parameter values, we avoided this strong assumption.

The computation of $\boldsymbol{I}_o$, which requires the joint density of (T,δ), not including Y, is facilitated by the use of the supplemented EM-algorithm (Meng and Rubin, 1991). The EM-algorithm implicitly defines a mapping $\boldsymbol{\Theta} \to \boldsymbol{M}(\boldsymbol{\Theta}) = (M_1(\boldsymbol{\Theta}),\ldots,M_d(\boldsymbol{\Theta}))'$ from the parameter space to itself such that $\widehat{\boldsymbol{\Theta}}^{(m+1)} = \boldsymbol{M}(\widehat{\boldsymbol{\Theta}}^{(m)})$ for $m = 0,1,\ldots$. A Taylor series expansion in the neighborhood of $\widehat{\boldsymbol{\Theta}}$ yields that

$$(\widehat{\boldsymbol{\Theta}}^{(m+1)} - \widehat{\boldsymbol{\Theta}})' \approx (\widehat{\boldsymbol{\Theta}}^{(m)} - \widehat{\boldsymbol{\Theta}})'\boldsymbol{DM}, \text{ where } \boldsymbol{DM} = \left(\frac{\partial M_j(\boldsymbol{\Theta})}{\partial\boldsymbol{\Theta}_i}\right)\Bigg|_{\boldsymbol{\Theta}=\widehat{\boldsymbol{\Theta}}},$$

a $d\times d$-matrix evaluated at $\boldsymbol{\Theta} = \widehat{\boldsymbol{\Theta}}$. Meng and Rubin (1991) further show that $\boldsymbol{I}_o^{-1} = \boldsymbol{I}_{oc}^{-1}(\boldsymbol{I}_d - \boldsymbol{DM})^{-1}$, with $\boldsymbol{I}_d$ a $d\times d$ identity matrix. For more details on the computation of $\boldsymbol{DM}$, we refer to chapter 12 of Gelman et al. (2004) and section 3.3 of Meng and Rubin (1991). Using (5), this leads to the AICcd as we use it in this paper,

$$\begin{aligned}\text{AICcd} &= -2Q(\widehat{\boldsymbol{\Theta}} \mid \widehat{\boldsymbol{\Theta}}) + 2\,\text{trace}(\boldsymbol{I}_d - \boldsymbol{DM})^{-1} \\ &= -2Q(\widehat{\boldsymbol{\Theta}} \mid \widehat{\boldsymbol{\Theta}}) + 2d + 2\,\text{trace}\{\boldsymbol{DM}(\boldsymbol{I}_d - \boldsymbol{DM})^{-1}\}. \end{aligned}\tag{6}$$

This criterion differs in two aspects from the infeasible AIC in (3). First, the expected complete data likelihood is used, and second, there is a correction to the penalty term

that takes the complexity of the modeling process due to the missing information into account. When all data are observed, $\boldsymbol{DM} = 0$ and the penalty reduces to the classical one.

We wish to mention that the mixture regression criterion of Naik et al. (2007) as an extension of the AIC to select both the number of components in the mixture and the variables within each component is not suitable for our purpose. Indeed, we know exactly the number of components in the mixture from the problem content, moreover even partial cluster membership is known. Only for censored observations the group membership is unknown. In addition, the mixture regression criterion assumes fully observed cases, while these data here are intrinsically censored.

2.3 AIC explicitly incorporating censoring

An alternative treatment of the censored observations is to treat the censored times as "missing" event times. The model that we wish to find should be well for describing the true event times U, and not only for the observed times T. Therefore, we start by writing the joint log likelihood of (T, U, Y) as, with $\boldsymbol{\Theta} = (\boldsymbol{\beta}, \boldsymbol{b})$,

$$L_n(\boldsymbol{\Theta}; T, U, Y) = \sum_{i=1}^{n} \Big\{ \log P_{Y_i}(\boldsymbol{z}_i; \boldsymbol{b}) + \log \tilde{f}_{Y_i}(T_i, U_i; \boldsymbol{\beta}) \Big\},$$

where $P_{Y_i}(\boldsymbol{z}_i) = \pi(\boldsymbol{z}_i; \boldsymbol{b})$ when $Y_i = 1$ and $P_{Y_i}(\boldsymbol{z}_i) = 1 - \pi(\boldsymbol{z}_i; \boldsymbol{b})$ when $Y_i = 0$. We define $\tilde{f}_{Y_i}(t_i; \boldsymbol{\Theta}) = f(t_i \mid Y_i = 1, \boldsymbol{\beta})^{\delta_i} S(t_i \mid Y_i = 1, \boldsymbol{\beta})^{(1-\delta_i)}$ when $Y_i = 1$ and take $\tilde{f}_{Y_i}(t_i) = 1$ when $Y_i = 0$. The Q-function for use in the EM-algorithm and the AIC can here be defined as,

$$\begin{aligned} Q(\boldsymbol{\Theta}_2, \boldsymbol{\Theta}_1) \;=\; & \sum_{i=1}^{n} \Big(\log \pi(\boldsymbol{z}_i, \boldsymbol{b}_2) + \mathrm{E}_f \big[\log \big\{ \tilde{f}_{Y_i}(T_i \mid Y_i = 1, \boldsymbol{\Theta}_2) \big\} \mid T_i, \boldsymbol{\Theta}_1 \big] \Big) w_{1i}(\boldsymbol{\Theta}_1) \\ & + \sum_{i=1}^{n} \Big(\log \left(1 - \pi(\boldsymbol{z}_i, \boldsymbol{b}_2)\right) + \log \big\{1\big\} \Big) (1 - w_{1i}(\boldsymbol{\Theta}_1)), \end{aligned}$$

where $w_{1i}(\boldsymbol{\Theta}_1) = P(Y_i = 1 \mid T_i = t; \boldsymbol{\Theta}_1)$ and the expectation 'E_f' is here computed with respect the model density of T, given $Y = 1$ and using parameter value $\boldsymbol{\Theta}_1$. With C_i the

censoring time for observation i, if $T_i \leq C_i$, the true event time is observed and $U_i = T_i$, while if $T_i > C_i$, the true event time U_i is not observed. Then we have that

$$\begin{aligned}
&\mathrm{E}_f\big[\log\big\{\tilde{f}_{Y_i}(T_i \mid Y_i = 1; \boldsymbol{\Theta}_2)\big\} \mid T_i, \boldsymbol{\Theta}_1\big] \\
&= \sum_{i=1}^{n} \delta_i \log f_{U_i|Y_i}(T_i \mid Y_i = 1, \boldsymbol{\Theta}_2) + \sum_{i=1}^{n}(1-\delta_i)\mathrm{E}_f[\log f_{U_i|Y_i}(U_i \mid Y_i = 1, \boldsymbol{\Theta}_2) \mid T_i, \boldsymbol{\Theta}_1].
\end{aligned}$$

This leads to defining the function Q for use in an EM-algorithm in the following way,

$$\begin{aligned}
Q(\boldsymbol{\Theta}_2 \mid \boldsymbol{\Theta}_1) &= \sum_{i=1}^{n} \log \pi(\boldsymbol{z}_i; \boldsymbol{b}_2) w_{1i}(\boldsymbol{b}_1, \boldsymbol{\beta}_1) + \sum_{i=1}^{n} \log(1 - \pi(\boldsymbol{z}_i; \boldsymbol{b}_2))\{1 - w_{1i}(\boldsymbol{b}_1, \boldsymbol{\beta}_1)\} \\
&+ \sum_{i=1}^{n} \delta_i \log f(T_i \mid Y_i = 1, \boldsymbol{\beta}_2) w_{1i}(\boldsymbol{b}_1, \boldsymbol{\beta}_1) \\
&+ \sum_{i=1}^{n} (1-\delta_i) \frac{\int_{c_i}^{\infty} \log f(u_i \mid Y_i = 1, \boldsymbol{\beta}_2) f(u_i \mid Y_i = 1, \boldsymbol{\beta}_1) du_i}{P(T_i \geq C_i \mid Y_i = 1, \boldsymbol{\beta}_1)} w_{1i}(\boldsymbol{b}_1, \boldsymbol{\beta}_1),
\end{aligned}$$

with

$$w_{1i}(\boldsymbol{\Theta}) = P(Y_i = 1 \mid T_i = t; \boldsymbol{\Theta}) = \begin{cases} \dfrac{\pi(\boldsymbol{z}_i, \boldsymbol{b}) f(t_i; \boldsymbol{\beta})}{\pi(\boldsymbol{z}_i, \boldsymbol{b}) f(t_i; \boldsymbol{\beta}) + (1 - \pi(\boldsymbol{z}_i, \boldsymbol{b}))} & \text{for } \delta_i = 0 \\ 1 & \text{for } \delta_i = 1. \end{cases}$$

Defining the AICcd proceeds as in Section 2.2 using this function Q. The resulting AICcd has a correct Kullback-Leibler interpretation for right-censored data from a mixture distribution. This way of incorporating the censoring provides (in models without mixture) an alternative to the AIC proposed by Suzukawa et al. (2001).

3 AIC for multiple event mixture cure models

We extend the parametric competing risk model of Watkins et al. (2013) by allowing for the semiparametric Cox proportional hazard model. In this model one distinguishes multiple events (e.g., default, prepayment) for which the time to event is important and considers another class of events (such as maturity) which happen at a fixed time. This class encompasses the group of 'immunes' in Section 2. Only those events for which

neither event takes place, are considered censored. For the formulation of this model, three indicators are used:

(1) Y_m, indicating that the loan is considered to be mature, so repayed at the indicated end date of the loan;

(2) Y_d, indicating that default takes place;

(3) Y_p, indicating that early repayment takes place.

Note that this set of (Y_m, Y_d, Y_p) is exhaustive and mutually exclusive. However, when an observation is censored, it is not known which event type will occur. In analogy to the equations (1) and (2), the unconditional survival function can be written as

$$\begin{aligned} S(t \mid \boldsymbol{x}_p, \boldsymbol{x}_d, \boldsymbol{z}) &= \pi_p(\boldsymbol{z}) S_p(t \mid Y_p = 1, \boldsymbol{x}_p) \\ &\quad + \pi_d(\boldsymbol{z}) S_d(t \mid Y_d = 1, \boldsymbol{x}_d) + \big(1 - \pi_p(\boldsymbol{z}) - \pi_d(\boldsymbol{z})\big), \end{aligned}$$

with S_p and S_d denoting the survival functions for, respectively, prepayment and default. Using the subscript '1' for default (d) and '2' for prepayment (p), the corresponding observed likelihood is given by

$$\begin{aligned} L_{\text{obs}}(\boldsymbol{\Theta}) &= \prod_{i=1}^{n} \Big\{ \prod_{j=1}^{2} \big(\pi_j(\boldsymbol{z}_i; \boldsymbol{b}_j) f_j(t_i \mid Y_{j,i} = 1, \boldsymbol{x}_{j,i}; \boldsymbol{\beta}_j)\big)^{Y_{j,i}} \big(1 - \sum_{j=1}^{2} \pi_j(\boldsymbol{z}_i; \boldsymbol{b}_j)\big)^{Y_{m,i}} \Big\}^{\delta_i} \\ &\quad \times \Big\{ \big(1 - \sum_{j=1}^{2} \pi_j(\boldsymbol{z}_i; \boldsymbol{b}_j)\big) + \sum_{j=1}^{2} \pi_j(\boldsymbol{z}_i; \boldsymbol{b}_j) S_j(t_{j,i} \mid Y_{j,i} = 1, \boldsymbol{x}_{j,i}; \boldsymbol{\beta}_j) \Big\}^{1-\delta_i}, \end{aligned}$$

where $\boldsymbol{\Theta} = (\boldsymbol{b}_p, \boldsymbol{b}_d, \boldsymbol{\beta}_p, \boldsymbol{\beta}_d)$. Note the flexibility of this model; each model part may employ its own set of covariates, hence the vectors $\boldsymbol{x}_d, \boldsymbol{x}_p$ and $\boldsymbol{z}$ may be different. We rewrite this model for use in an EM algorithm such that the AICcd of (6) may be applied for model selection. For this purpose, we start from the complete likelihood, hence the likelihood expression under the assumption that full information on $\boldsymbol{Y} = (Y_m, Y_d, Y_p)$ is present

$$L_{\text{complete}}(\boldsymbol{\Theta}; \delta_i, Y_i, T_i) = \prod_{i=1}^{n} \Big\{ \prod_{j=1}^{2} \big(\pi_j(\boldsymbol{z}_i; \boldsymbol{b}_j)\big)^{Y_{j,i}} \big(1 - \prod_{j=1}^{2} \pi_j(\boldsymbol{z}_i; \boldsymbol{b}_j)\big)^{Y_{m,i}} \Big\}$$

$$\times\Big\{\prod_{j=1}^{2}\big(h_j(t \mid Y_{j,i}=1,\boldsymbol{x}_{j,i};\boldsymbol{\beta}_j)^{\delta_i}S_d(t_{j,i} \mid Y_{j,i}=1,\boldsymbol{x}_{j,i};\boldsymbol{\beta}_j)\big)^{Y_{j,i}}\Big\}$$

Converting to the log likelihood and computing the expected value this time using the model density with parameter $\boldsymbol{\Theta}_1$ leads us to the Q-function as given in (4),

$$\begin{aligned}
Q(\boldsymbol{\Theta}_2 \mid \boldsymbol{\Theta}_1) &= \mathrm{E}_f[\log L_{\text{complete}}(\boldsymbol{\Theta_2};T_i,\delta_i,Y_i) \mid T_i,\delta_i,\boldsymbol{\Theta_1}] \\
&= \sum_{i=1}^{n}\Big\{\sum_{j=1}^{2} w_{ji}\log(\pi_j(\boldsymbol{z}_i;\boldsymbol{b}_j)) + w_{mi}\log(1-\sum_{j=1}^{2}\pi_j(\boldsymbol{z}_i;\boldsymbol{b}_j)) \\
&\quad + \sum_{j=1}^{2} w_{ji}\delta_i\log(h_j(t_i \mid Y_j=1,\boldsymbol{x}_{j,i};\boldsymbol{\beta}_j)) + w_{ji}\log(S_j(t_i \mid Y_j=1,\boldsymbol{x}_{j,i};\boldsymbol{\beta}_j))\Big\}.
\end{aligned}$$

Note that conditional expectations of $Y_{j,i}$ $(j=1,2)$, $\mathrm{E}_f\big[Y_{j,i} \mid T_i,\delta_i,\boldsymbol{\Theta}_1\big]$, are computed here with respect to the model density using parameter $\boldsymbol{\Theta}_1$ and are denoted by w_{ji} with $w_{mi}=1-w_{1i}-w_{2i}$ and for $j=1,2$,

$$\begin{aligned}
w_{ji} &= w_{ji}(\boldsymbol{\Theta}) = P(Y_{j,i}=1 \mid T_i=t;\delta_i;\boldsymbol{\Theta}) \\
&= \begin{cases}
\dfrac{\pi_j(\boldsymbol{z}_i,\boldsymbol{b}_j)S_j(t_i;\boldsymbol{\beta}_j)}{\sum_{k=1}^{2}\pi_k(\boldsymbol{z}_i;\boldsymbol{b}_k)S_k(t_i;\boldsymbol{\beta}_k)+(1-\sum_{k=1}^{2}\pi_k(\boldsymbol{z}_i;\boldsymbol{b}_k))} & \text{for } \delta_i=0 \\
1 & \text{for } Y_{j,i}=1 \text{ and } \delta_i=1 \\
0 & \text{for } Y_{j,i}=0 \text{ and } \delta_i=1.
\end{cases}
\end{aligned}$$

4 Simulation study

4.1 Simulation settings

All computations were performed in R, adapting the library `smcure` (Cai et al., 2012) to produce the AICcd values.

Three different simulation settings were used. For each simulation setting, 100 simulation runs with n=5000 observations and 5 variables were executed. The probability of being susceptible, that is $(1-\pi(\boldsymbol{z}))$ was generated using the relationship $\pi(\boldsymbol{z}) = \dfrac{\exp(\boldsymbol{b}'\boldsymbol{z})}{1+\exp(\boldsymbol{b}'\boldsymbol{z})}$, with variables z_1–z_5 of which the distributions are stated in Table 1 and

variable	z_1	z_2	z_3	z_4	z_5
Distr.	$\text{Bin}(n, 0.7)$	$\Gamma(\lambda = 2.74, r = 1.3086)$	$N(1,1)$	$N(1,2)$	$\text{Bin}(n, 0.66)$

Table 1: Distributions of z_1–z_5 used in the simulation study.

parameter	(intercept)	b_1	b_2	b_3	b_4	b_5	β_1	β_2	β_3	β_4	β_5
Setting 1 & 3	3	3.5	-1	0	0	-1	2.5	0	0	-1	0
Setting 2	1	1.5	-1.5	0	0	-1.8	2.5	0	0	-1	0

Table 2: Simulation study. Parameter values of the true model.

with parameters $\boldsymbol{b}$ as in Table 2. True Y-values are consequently generated through a sample vector of 0 and 1 using these probabilities $\pi(\boldsymbol{z})$. For the uncured part of the population, Weibull default times (shape parameter = 1, scale parameter=0.5) were generated, using the same five variables (thus $\boldsymbol{x} = \boldsymbol{z}$) with the distributions and parameter values $\boldsymbol{\beta}$ as in Tables 1 and 2. For the first two simulation settings, censoring times were uniformly distributed on the interval $[0, 1]$. For setting 3, censoring times were uniformly distributed on the interval $[0, 20]$, in order to lower the amount of censoring compared to settings 1 and 2. Each time we performed an exhaustive model search, thus $(2^5 - 1)^2 = 961$ AICcd's were calculated for every simulation run.

In the first simulation setting, the censoring percentage was 60% (hence, around 3000 observations were censored, $\delta = 0$) and 80% of the observations were susceptible ($Y = 1$). For setting 2, we mimicked the situation of the data example (see Section 5), resulting in the uncensored percentage nearly equal to 10%, and the susceptible percentage of the observations equal to 20%. For setting 3 the censoring time interval was increased from [0,1] to [0,20], resulting in more observed defaults, and less censoring. Because of this, the real default time was observed for 70% of the observations, with 80% susceptible

observations as in setting 1.

For comparison purposes, four other versions of AIC were calculated.

$$\mathrm{AIC_{cs}} = -2\log L_{\mathrm{Cox}}\left(\hat{\boldsymbol{\beta}};\boldsymbol{x}\right) + 2d_{\mathrm{Cox}}\ , \quad \mathrm{AIC_{cl}} = -2\log L_{\mathrm{Cox}}\left(\hat{\boldsymbol{\beta}};\boldsymbol{x}\right) + 2d_{\mathrm{Cox,Log}}\ ,$$
$$\mathrm{AIC_{ls}} = -2\log L_{\mathrm{Log}}\left(\hat{\boldsymbol{b}};\boldsymbol{z}\right) + 2d_{\mathrm{Log}}\ , \quad \mathrm{AIC_{ll}} = -2\log L_{\mathrm{Log}}\left(\hat{\boldsymbol{b}};\boldsymbol{z}\right) + 2d_{\mathrm{Cox,Log}}.$$

The first subscripts of the AIC's is either c or l, which stands for "Cox" or "Log" and indicates the likelihood of the survival or logistic part of the mixture only. The second subscript indicates whether a "short" (s) or "long" (l) penalty term was used. A short penalty term means that the parameters accounted for are only calculated by the model specified in the first subscript, and a long penalty term incorporates all the parameters. The penalty is defined to be twice the number of considered parameters.

The reason for comparing the AICcd to those at first sight rather naive AIC-calculations, is because in practice, those AICs might by some researchers be in use instead of the corrected version with complete-data log likelihoods when analyzing mixture cure models. We want to investigate whether it is reasonable to use those AICs. We are not aware of other model selection criteria for mixture cure models.

4.2 Simulation Results

Table 3 summarises some model selection aspects of the AICs. The results of all simulation runs were averaged. Next to the type of AIC used, we list the ranking (among the 961 models) of the true model as simulated. The next four columns indicates the average number of variables that were lacking in the selected model (-) or were unnecessarily included in the selected model (+) for the log-component and the Cox-component respectively as compared to the "true" model. The last two columns are the joint averaged over- and underselection values.

The simulated data were generated using three true variables for the log-model, and

Sett.	Method	Mean rank	Log -	Log +	Cox -	Cox +	Total -	Total +
MAX		961	3	2	2	3	5	5
1	AICcd	107.85	1.14	0.90	0.02	1.68	1.16	2.58
	AICcs	163.13	1.63	1.51	0.00	0.66	1.63	2.17
	AICcl	163.91	1.70	1.19	0.00	0.67	1.70	1.86
	AICls	155.26	1.43	0.44	0.67	1.58	2.10	2.02
	AICll	151.67	1.46	0.48	0.67	1.13	2.13	1.61
2	AICcd	59.81	0.00	1.32	0.17	1.44	0.17	2.76
	AICcs	95.06	0.88	1.51	0.01	0.83	0.89	2.34
	AICcl	94.95	1.02	1.07	0.01	0.76	1.03	1.83
	AICls	162.64	0.02	1.46	1.99	1.55	2.01	3.01
	AICll	159.58	0.02	1.43	2.00	1.17	2.02	2.60
3	AICcd	13.01	0.00	0.84	0.00	0.91	0.00	1.75
	AICcs	79.53	1.28	1.49	0.00	0.41	1.28	1.90
	AICcl	80.39	1.35	0.99	0.00	0.41	1.35	1.40
	AICls	151.68	2.58	1.16	1.06	2.32	3.64	3.48
	AICll	147.33	2.59	1.17	1.06	2.02	3.65	3.19

Table 3: Simulation settings 1 – 3, 100 runs for an exhaustive search. Averages for underfitting (-) and overfitting (+) in terms of variables as compared to the true model, for each part of the mixture model, and for the combined parts (total).

two variables for the Cox-model. The first line in Table 3 indicates the maximum value possible for each column of the table. A perfect selection would give a mean rank of 1 (= the "true" model is always selected), and 0-values for all the other entries, indicating that all necessary variables are present in the model, and all the unnecessary variables are left out. AIC is known to be an efficient model selection method with regard to mean squared

prediction error (Claeskens and Hjort, 2008, Chap 4), though not to be consistent hence we do not expect to find small ranks for the true model here. Moreover, the chosen settings are quite demanding with large percentages of censored data (especially for settings 1 and 2), which are typical to credit risk studies, as opposed to medical studies where those percentages are usually much smaller.

The simulation study indicates that for these settings the Cox part of the log-likelihood is dominant, both in magnitude and for model selection purposes. In Table 3, we see that AICcd outperforms the other criteria regarding the mean rank of the true model for all three settings. Overfitting proportions are favorable for the low-censored setting (Setting 3), but quite high for Setting 1 and 2. On the other hand, underfitting proportions are low for the AICcd compared to the other measures. This is an important result as underfitting (missing important predictors) is considered worse than overfitting. When looking at the change in result as the censoring percentage changes, it becomes clear that high percentages of censored cases on one hand (setting 2) and a big discrepancy between observed versus true defaults (setting 1) have a negative impact on the performance of any information criterion. This gives us a strong indication that it would be advisable to incorporate additional information (such as in the multiple event models) to reduce the number of censored cases.

A comparison with the simpler criterion that just counts the number of parameters is for the chosen settings not behaving too badly, since it turns out that the correction term involving $\boldsymbol{DM}$ takes values in a bounded range, and is here not influencing the model order too much. Again, we stress that no other information criteria have yet been developed for these mixture models, which could have made the comparison more interesting. For comparisons of AICcd in regression models to other AIC-like versions we refer to Cavanaugh and Shumway (1998).

5 Variable selection for a credit loan dataset

5.1 Data and method

The survival analysis techniques were applied to personal loan data from a major U.K. financial institution. All customers are U.K. borrowers who had applied to the bank for a loan. The data set consisted of the application information of 50,000 personal loans, together with the repayment status for each month of the observation period of 36 months. We note that the same data were also used in Stepanova and Thomas (2002) and later by Tong et al. (2012). In this paper only a subset of the loans with loan term 36 months were used for the analysis (containing n =7521 observations).

An account was considered as a default (censoring indicator=1) if it was at least 90 days in arrears. When an account was not in arrears or only in arrears for less than 90 days, the account was considered as a non-default (censoring indicator=0). As for most credit data, the percentage of defaults within the observation period was very low: default was only observed for 376 of the 7521 observations. In Section 5.3 we reconsider this dataset taking prepayments and maturity into account, hereby reducing the number of censored cases.

For each observation, we considered 8 candidate covariates, see Table 4. In the model selection approach of Section 5.2, we searched through all subsets of the collection of 8 covariates, and this for both model components, resulting in $(2^8 - 1)^2 = 65025$ AICcd values, where we have excluded empty latency and incidence models. Using the same method of exhaustive search for the modelling approach in Section 5.3 would result in over 16 581 375 AICcd calculations ($(2^8 - 1)^3$), because this time three different covariate vectors are considered. Therefore, instead of an exhaustive search, a genetic algorithm was used to find a good model, for which we used the package `GA` in R (Scrucca, 2013). We used this package with AICcd in the binary representation indicating the presence (1)

	Description	Type
v_1	The gender of the customer (1=M, 0=F)	categorical
v_2	Amount of the loan	continuous
v_3	Number of years at current address	continuous
v_4	Number of years at current employer	continuous
v_5	Amount of insurance premium	continuous
v_6	Home phone or not (1=N,0=Y)	categorical
v_7	Own house or not (1=N, 0=Y)	categorical
v_8	Frequency of payment(1=low/unknown, 0=high)	categorical

Table 4: Credit loan data. Description of the variables.

or absence (0) of a specific variable, and with all default settings, i.e., population size 50, crossover probability 0.8, mutation probability 0.1. For the model selection purpose, this algorithm starts with randomly including and excluding some variables. The algorithm consists of several "generations", and at the end of each generation, the AICcd-values of the inspected models are evaluated, and the models with the lowest AICcd-values are withheld in the next generations. Starting from those models, small changes are made with the purpose to find models with even lower AICcd-values.

5.2 Variable selection for the time to default

After calculating the AICcd values for each of the considered models, the models were sorted according to their resulting AICcd values. Seven models will be discussed and compared: the best five models according to the AICcd, the full model and (again according to AICcd) the best model under the restriction that the latency and incidence model should contain the same covariates; see Table 5.

Model	AICcd	Rank	Part	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8
Best	7372.85	1	Incidence	1	0	1	1	1	1	1	1
			Latency	1	0	0	1	1	1	1	0
Second best	7373.06	2	Incidence	1	1	1	1	1	1	1	1
			Latency	1	0	0	1	1	1	1	0
Third best	7385.11	3	Incidence	1	1	0	1	1	1	1	1
			Latency	1	0	0	1	1	1	1	0
Fourth best	7385.28	4	Incidence	1	1	1	1	1	1	1	0
			Latency	1	0	0	1	1	1	1	0
Fifth best	7385.79	5	Incidence	1	0	1	1	1	1	1	0
			Latency	1	0	0	1	1	1	1	0
Full	7446.92	215	(both)	1	1	1	1	1	1	1	1
Same covar.	7397.87	17	(both)	1	0	0	1	1	1	1	0

Table 5: Credit loan data. Variables contained in the five best models according to AICcd, the full model and the AICcd-best model with the same parameters in both model parts. The value of AICcd, as well as its ranking is given.

We observe that for all the five best models, the same latency model is selected whereas the incidence model covariates vary. For this dataset, the incidence model seems to require more variables. Whereas variables v_2 (amount of the loan), v_3 (number of years living at a current address) and v_8 (frequency of the payment) are never included in the latency part of the best five models, those three variables are also the ones left out in the incidence model, but at the most with two at the same time. The full model only ranks 215th with regard to AICcd value. The same covariate model, for this dataset, uses the same covariates as for the latency part. Its rank is 17, with a difference in AICcd values as

Month	Best	Second best	Third best	Full	Same covar.
18	0.710	0.709	0.695	0.707	0.703
24	0.700	0.700	0.683	0.700	0.688
36	0.688	0.685	0.664	0.684	0.671

Table 6: Credit loan data. AUC values for the top three models according to AICcd, the full model and the AICcd-best model with the same variables in both model parts, when predicting default at 18, 24 and 36 months respectively.

compared to the best model equal to about 25, clearly showing the preference for the separate covariate parts.

In the credit risk context, a widely used method to evaluate binary classifiers is by means of the receiver operating characteristics curves. These curves give the percentage of correctly classified observations for each possible threshold value. The specific measure of interest is the area under the curve (AUC), which can also be used in the context of survival analysis (Heagerty and Saha, 2000). We computed the AUC values for 5 models of interest, when predicting default at three different time instances (18, 24 and 36 months). Each time, 2/3 of the data was used as a training set, and 1/3 as a test set. The AUC-values can be found in Table 6.

In Table 7, the parameter estimates of the best model according to AICcd can be found. Positive $\boldsymbol{b}$-parameters have a positive impact on the probability of being susceptible, and positive $\boldsymbol{\beta}$-parameters shorten the time until default. As a result, working at the same employer for a longer time period decreases the risk to default, as well as having a home phone and owning a house (binary variables decoded as 1 = no and 2 = yes). The gender of a subject has an ambiguous effect on default: whereas being male lowers the probability of being susceptible, we see that the time until default when susceptible is shorter for men.

Figure 1 presents the estimated survival curves for two randomly chosen persons in

Part	Int.	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8
Inc. ($\hat{\boldsymbol{b}}$)	-1.586	-0.311	–	-0.036	-0.044	0.001	0.002	0.328	-0.380
(se)	(0.210)	(0.155)	–	(0.009)	(0.014)	(0.0002)	(0.285)	(0.129)	(0.120)
Lat. ($\hat{\boldsymbol{\beta}}$)	–	0.551	–	–	-0.066	0.0003	0.852	0.024	–
(se)	–	(0.177)	–	–	(0.019)	(0.0002)	(0.304)	(0.172)	–

Table 7: Credit loan data. The parameter estimates for the time to default with their standard errors (se) for the AICcd-best model for the incidence (Inc.) and latency (Lat.) parts of the model. Variables not selected were not estimated.

the dataset (namely a male person, not possessing a home phone and working at the same employer for a relatively short time, and a female person, possessing a home phone and working at the same employer for a relatively long time). We consider estimates obtained in the best mixture cure model with different covariates for both model parts, in the best such model with the same covariates, and in the best Cox proportional hazard model with all variables except for the customer's gender. This was the model selected by the AIC using the partial likelihood and penalizing for the number of parameters in the model.

For the male person, the estimated survival percentages were relatively high, and all three approaches give reasonably close estimates. However, for the female person with lower values for the estimated survival probabilities, we observe a clear difference with the estimates from the mixture model and with that of the Cox proportional hazard model. The estimated proportion in the mixture was equal to 12.81 % for this subject, clearly suggesting the need of the mixture model. For this data example, the use of the same covariates leads to larger estimated probabilities for survival.

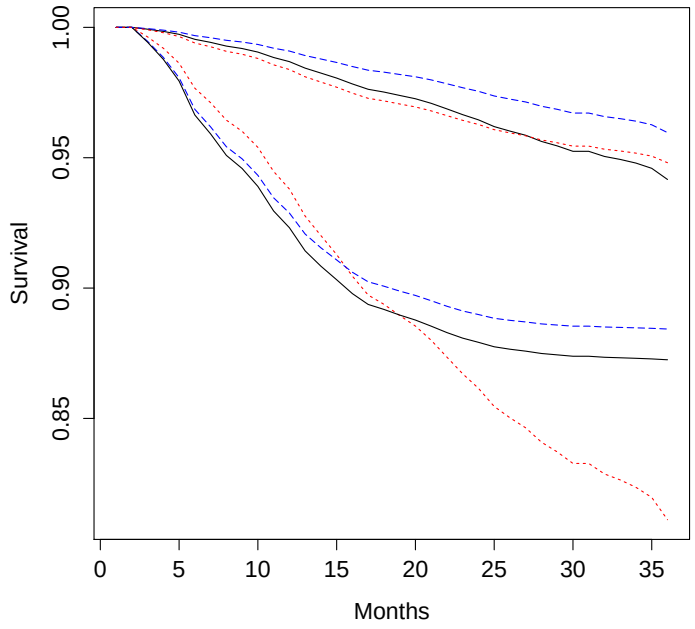

Figure 1: Credit loan data. Estimated survival curves for two observations using three models. In solid line type (black) we show the estimates for the selected best model, the dashed lines (blue) use the same-covariate best model, while the dotted lines (red) give the estimated survival curve using the Cox proportional hazard model, ignoring the mixture.

5.3 Variable selection for the multiple event model

As stated before, the multiple event model does not only incorporate default, but also early repayment, resulting in two incidence models and two latency models. For this dataset there are 3.6% observations (269 cases) for which maturity has occurred (so, which are belonging to the "cured" fraction), 5% (376 cases) were in default, and 39.8% (2992 cases) have prepayments. The remaining 51.6% are truly censored observations.

The genetic algorithm used is part of the package `GA` in R by Scrucca (2013), with default settings, as described in section 5.1. Despite the fact that genetic algorithms are

quite successful and efficient, it is never certain that the final outcome will yield the overall lowest AICcd value. However, the genetic algorithm we used was also applied to the data example for the mixture cure model in section 5.2, resulting in precisely the same selected model as with the exhaustive search. The resulting model for the joint analysis of default and prepayment with parameter estimates can be found in Table 8. The interpretation of the parameters in Table 8 is similar to the mixture cure-interpretation. Again, we see that not having a home phone increases the probability and shortens the time for default (both positive $\hat{\boldsymbol{b}}$- and $\hat{\boldsymbol{\beta}}$ values). A longer working duration at the same employer, however, decreases the probability of default but has no significant result on the time until default according to the model selected by the genetic algorithm using AICcd. The number of parameters included in the latency model of default has gone from five parameters in the mixture cure model to four parameters in the multiple event incidence model. A possible explanation is that since more information is gained by adding an early repayment part, less predictors are needed for the time until default. For the early repayment parameters, we notice that five variables are included in the latency part. We see that male subjects tend to have a lower chance to belong to the early repayment group ($\boldsymbol{b} < 0$), but when belonging to that group, they tend to prepay earlier than female subjects. Note that the same variables are included for the two incidence models, where only v_7 is not in the incidence model. This is a result of the fact that the respective probabilities are estimated in one multinomial logit model (as we have now three groups: early repayment, default and maturity). The sign of $\hat{\boldsymbol{b}}_{\boldsymbol{d}}$ and $\hat{\boldsymbol{b}}_{\boldsymbol{p}}$ gives the relation between default and early repayment respectively, in relation to maturity. For example: the multinomial log-odds for a certain subject to belong to the early repayment-group versus the mature group are expected to increase by 0.083 units (ceteris paribus) when the subject does not have a homephone, however, the log-odds to belong to the default-group compared to the mature group are even more elevated (increase by 0.481 units).

Part	Intercept	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8
$\hat{\boldsymbol{b}}_d$	-0.837	-0.084	-0.00007	-0.038	-0.094	0.001	0.481	–	-0.479
$\hat{\boldsymbol{\beta}}_d$	–	0.118	–	–	–	0.0001	0.342	–	0.106
$\hat{\boldsymbol{b}}_p$	0.648	-0.174	-0.00001	-0.020	-0.014	-0.00001	0.083	–	-0.084
$\hat{\boldsymbol{\beta}}_p$	–	0.073	-0.00003	–	–	–	-0.359	-0.081	0.163

Table 8: Credit loan data. The parameter estimates for the multiple event incidence model as found by the genetic algorithm.

As a final illustration, the default and early repayment curves were plotted in Figure 2 for the same two random observations as for the mixture cure model. The male person incurs a higher risk regarding default, and a lower propensity regarding early repayment.

6 Discussion

The development of advanced survival models for credit risk data is in current progress. With this paper we contributed with the derivation of a proper variable selection method. We have used the popular Akaike information criterion as the basis of the selection procedure. By making use of the output of the EM procedure for model fitting, we obtained a relatively simple criterion and have implemented this procedure in R, making use of existing packages for fitting mixture cure models.

Emphasizing other aspects of the modeling procedure would lead to the development of other selection methods. A Bayesian information criterion for these models is expected to have consistency properties, however, under the strong (and unrealistic) assumption that the true credit risk model is exactly described by one of the used models. A focused information criterion (Claeskens and Hjort, 2003) would rather assume local misspecification and selects a model that is best in terms of mean squared error or mean squared

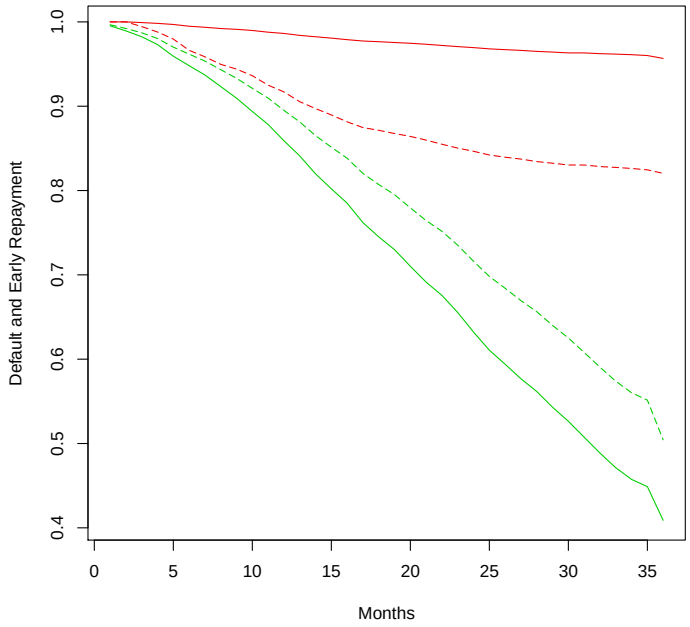

Figure 2: Credit loan data. Estimated probabilities for default and early repayment for two observations. The green (steeper) lines represent early repayment, and the flatter lines default. The solid line represents a female person, possessing a home phone and working at the same employer for a relatively long time, and the dashed lines a male person, not possessing a home phone and working at the same employer for a relatively short time.

prediction error for a certain focus quantity (such as the probability of the time to default to fall in a certain period). Some of these approaches are under current investigation.

Our simulation study and the data analysis have illustrated that using different covariate vectors may lead to better models regarding AUC value and regarding to model ranking according to AICcd. Not restricting to same-covariate models for mixture modeling is worthwhile, our variable selection approach easily allows for such general modeling strategies. The use of a genetic search algorithm in combination with the AICcd provides a handy way of incorporating many variables.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. and Csáki, F., editors, *Second International Symposium on Information Theory*, pages 267–281. Akadémiai Kiadó, Budapest.

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., and Vanthienen, J. (2003). Benchmarking state of the art classification algorithms for credit scoring. *Journal of the Operational Research Society*, 54(6):627–635.

Bellotti, T. and Crook, J. (2009). Credit scoring with macroeconomic variables using survival analysis. *Journal of the Operational Research Society*, 60:1699–1707.

Cai, C., Zou, Y., Peng, Y., and Zhang, J. (2012). smcure: An R-package for estimating semiparametric mixture cure models. *Computer Methods and Programs in Biomedicine*, 108:1255–1260.

Cavanaugh, J. E. and Shumway, R. H. (1998). An Akaike information criterion for model selection in the presence of incomplete data. *Journal of Statistical Planning and Inference*, 67(1):45–65.

Claeskens, G. and Consentino, F. (2008). Variable selection with incomplete covariate data. *Biometrics*, 64:1062–1069.

Claeskens, G. and Hjort, N. (2003). The focused information criterion. *Journal of the American Statistical Association*, 98:900–916. With discussion and a rejoinder by the authors.

Claeskens, G. and Hjort, N. (2008). *Model Selection and Model Averaging*. Cambridge University Press, Cambridge.

Donohue, M. C., Overholser, R., Xu, R., and Vaida, F. (2011). Conditional Akaike information under generalized linear and proportional hazards mixed models. *Biometrika*.

Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Annals of Statistics*, 30:74–99.

Farewell, V. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, pages 1041–1046.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian data analysis*. Texts in Statistical Science Series. Chapman & Hall/CRC, Boca Raton, FL, second edition.

Heagerty, P. and Saha, P. (2000). SurvivalROC: time-dependent roc curve estimation from censored survival data. *Biometrics*, 56(2):337–344.

Hjort, N. and Claeskens, G. (2006). Focussed information criteria and model averaging for Cox's hazard regression model. *Journal of the American Statistical Association*, 101:1449–1464.

Ibrahim, J. G., Zhu, H., and Tang, N. (2008). Model selection criteria for missing-data problems using the EM algorithm. *J. Amer. Statist. Assoc.*, 103(484):1648–1658.

Kuk, A. and Chen, C. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, 79(3):531–541.

Liang, H. and Zou, G. (2008). Improved AIC selection strategy for survival analysis. *Computational Statistics & Data Analysis*, 52(5):2538 – 2548.

Meng, X.-L. and Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association*, 86(416):899–909.

Naik, P. A., Shi, P., and Tsai, C.-L. (2007). Extending the Akaike information criterion to mixture regression models. *Journal of the American Statistical Association*, 102(477):244–254.

Scrucca, L. (2013). GA: A package for genetic algorithms in R. *Journal of Statistical Software*, 53(4):1–37.

Stepanova, M. and Thomas, L. (2002). Survival analysis methods for personal loan data. *Operations Research*, 50(2):277–289.

Suzukawa, A., Imai, H., and Sato, Y. (2001). Kullback-Leibler information consistent estimation for censored data. *Ann. Inst. Statist. Math.*, 53(2):262–276.

Sy, J. P. and Taylor, J. M. G. (2000). Estimation in a Cox proportional hazards cure model. *Biometrics*, 56(1):227–236.

Tong, E. N. C., Mues, C., and Thomas, L. C. (2012). Mixture cure models in credit scoring: if and when borrowers default. *European Journal of Operational Research*, 218(1):132–139.

Van Gestel, T. and Baesens, B. (2008). *Credit Risk Management : Basic Concepts: Financial Risk Components, Rating Analysis, Models, Economic and Regulatory Capital*. OUP Oxford.

Watkins, J. G. T., Vasnev, A. L., and Gerlach, R. (2013). Multiple event incidence and duration analysis for credit data incorporating non-stochastic loan maturity. *Journal of Applied Econometrics*. to appear.

Xu, R., Vaida, F., and Harrington, D. P. (2009). Using profile likelihood for semiparametric model selection with application to proportional hazards mixed models. *Statist. Sinica*, 19(2):819–842.